DOCUMENT RESUME

ED 242 761 TM 840 174

AUTHOR Haladyna, Thomas M.
TITLE Increasing Information from Multiple-Choice Test Items.
PUB DATE Apr 84
NOTE 20p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 24-26, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; *Cutting Scores; High Schools; *Multiple Choice Tests; Pass Fail Grading; *Predictive Validity; *Scoring; Test Items; *Test Reliability
IDENTIFIERS *Distractors (Tests); *Option Weighting

ABSTRACT

The purpose of this study is to examine an option-weighting method as it affects pass-fail decisions in formative and summative evaluation of student achievement for instructional units, certification, advancement, licensure, admissions, placement, and selection. A database was constructed using high school achievement test data where a unidimensional domain was established and short tests simulated to test the validity of the option weighting procedure against the standard procedure for scoring test items. The weight for each option was its discrimination index. It has long since been established that option weighting positively affects estimates of reliability but has unclear effects on predictive validity. In this study, reliability was directly computed, and option weighting again proved effective in increasing reliability. More importantly, option weighting applied to a pass-fail decision-making situation provided a substantial improvement over regular scoring for tests which were constructed by simple random sampling from a well-defined domain. The advantages for option weighting were greatest when the test contained 10 items; as test length increased, the advantage of option weighting seemed to lessen. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

INCREASING INFORMATION FROM MULTIPLE-CHOICE TEST ITEMS

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
T. Haladyna
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

THOMAS M. HALADYNA

THE AMERICAN COLLEGE TESTING PROGRAM

ED242791

Introduction

Traditionally, the multiple choice test item format is used in a variety of ways in standardized and teacher-made achievement tests. Some of these uses include formative and summative evaluation of student achievement for instructional units, certification, advancement, licensure, admissions, placement, and selection. Many of these activities require pass-fail decisions.

A common convention in the scoring of multiple-choice items is to assign one point for a correct response and no points for an incorrect response. For quite some time, a number of researchers on testing have experimented with ways to increase the information yield in multiple-choice testing through option weighting. This practice involves the assigning of different values for each option. For example, for a single four-option item, four different values could be assigned based on the particular option weighting procedure used.

The purpose of this study is examine an option-weighting method as it affects pass-fail decisions in any of the above named situations. As a background to study, some salient research on option weighting is presented and discussed in light of the study.

Research on Option Weighting

Option weighting was first suggested by Guttman (1941), and since that time there has been a continued optimism for its use. A number of alternate procedures for option weighting have been proposed. These procedures take two fundamental forms. The first of these involve the judgment of content experts

about the inherent values of incorrect options (Patnaik and Traub, 1973). The second of these involve statistical analyses of item responses. A number of procedures have been proposed over the years, and these are more comprehensively reviewed in Haladyna (1984). Briefly these procedures include:

1. Guttman weighting, using the total score as a criterion;

2. Guttman weighting, using another measure as a criterion;

3. Successive integers, which is a rank order weighting system--the most discriminating option gets the highest value, the least discriminating the lowest value, in rank order;

4. Biserial/point biserial correlation between item performance and total test performance;

5. Factor analytic (Serlin and Kaiser; 1978; Willson, 1982);

6. Multiserial correlation (Donlon and Fitzpatrick, 1978);

7. Relative uncertainty (Donlon and Fitzpatrick, 1978); and

8. Reciprocal averages, which appear to resemble Guttman's first procedure (Baker and Hoyt, 1972).

These procedures are computationally feasible due to computers and computer programs expressly designed to compute weights. Most procedures are highly interrelated, and offer little advantage over the others.

The research on option weighting has led to two indisputable conclusions:

1. Option weighting almost always improves test score reliability.

2. While reliability is typically improved, predictive validity will seldom improve.

The first conclusion is supported in a number of studies (Bayuk, 1974; Bejar and Weiss, 1977; Claudy, 1978; Cross and Fry, 1978; Davis, 1959; Davis and Fifer, 1959; Downey, 1974; Echternacht, 1976; Hendrickson, 1971; Raffeld,

1975; Reilly, 1975; Reilly and Jackson, 1975; Sabers and White, 1969; Waters, 1975). In these studies, option weighting has the affect of spreading out test scores and thereby improving the internal consistency estimate of reliability.

The second conclusion appears to be also based on a number of studies (Claudy, 1978; Downey, 1974; Echternacht, 1976; Hendrickson, 1971; Raffeld, 1975; Reilly, 1975; Reilly and Jackson, 1975; Serlin and Kaiser, 1978; Sabers and White, 1969; Waters, 1975). The phenomenon apprently present in these studies is the attenuation paradox, reliability increases without affecting predictive validity. Some authors (e.g. Hendrickson, 1971; Hendrickson and Green, 1975; Serlin and Kaiser, 1978) have speculated that option weighting changes the trait, purifies it, to the extent that what is being measured is more internally consistent and less related to the criterion. No studies reviewed here focused on content validity or an orientation toward domain-referenced test score interpretations, which are more commonplace in instruction where instructional domains are well defined and large pools of items exist representing these domains. This study was intended to examine the relationship between one option-weighting procedure and the consistency of pass-fail decisions.

Logical Rationale for Option Weighting

Any domain of knowledge represented by a set of multiple-choice test items, or any other types of items for that matter, can be viewed in countable units. The most convenient and meaningful scale is percentage correct. Any random sample of test items administered to a student provides a maximum liklihood, unbiased estimate of performance in that domain. Therefore a percentage correct score of 65% tells us that about 65% of all items in the domain can be correctly answered by that student. Following such an

observation, the test designers, instructor, or test user may wish to make some decisions about the examinee's status regarding further instruction, competence to practice, or advancement.

Despite the widespread use of the zero-one convention for scoring multiple-choice items, the research on option weighting coupled with the compelling logic of option weighting stands in favor of this practice. Fundamentally, each item represents that domain of knowledge. Assigning one or no points for a response stands opposed to the finer discrimination implied. Carroll (1961) calls this practice "censoring," and the net effect of it is to obscure information. We know that with every multiple- choice item, there is usually an indisputable right answer. Agreement by content experts is one of the best ways to identify a correct response, but the "votes" of examinees is another good confirmatory method. These votes take the form of the point-biserial correlation between the item and total test performance, the discrimination index. Items with good discrimination indexes are generally favorably reviewed and used in future examinations. Fortunately, all items have more than one discrimination index--if one considers the fact that performance on each option can be correlated with the total test score. Henrysson (1971) has observed that incorrect options are often designed to appeal to the lower scoring examinees, therefore it should not be surprising that incorrect options are often highly and negatively correlated with the total test score, or very plausible, second-best options have positive correlations with the total test score. It makes sense to assign discriminative values for all answers. A very poor response deserves one kind of value, while a second-best response deserves a higher value. A multiple-choice item could conceivably take the form and values as follows:

Donnie was planning to paint two walls in a room with eight foot ceilings. One wall was 10 feet long, the other was 12 feet long. He had one quart of paint which was good enough to cover 72 square feet. How much more paint does he need?

-.40	A.	one quart
+.56	B.	one and one-half quarts
+.22	C.	two quarts
+.13	D.	three quarts

Option A is not a good response because one quart is not enough paint. Option B is the correct response. Option C is not too far from the correct response. If Donnie purchased two quarts, he could could paint the wall, but he would have some left over paint. Option D is too much paint, and while it is wrong it still is somewhat correct because it is also enough paint. Thus a logical analysis might lead to option weights just as a statistical analysis might accomplish the same.

If the items were given to a large group of students, and item analysis conducted, discrimination indexes might occur as those shown next to the options. It should be clear from this example that option weighting may have a logical rationale as well.

Method

Design of the Study

To address the issue posed in the previous section regarding pass-fail decisions, a data base was established for the study of the item-weighting procedure. Several tenets are presented here as each is salient to the study. Each of these tenets has firm foundations in classical test theory (Lord and Novick, 1968; Nunnally, 1978), generalizability theory (Brennan, 1983), item response theory (Lord and Novick, 1968), and criterion-referenced testing literature (Hambleton, Swaminanthan, Algina and Coulsen, 1978; Roid and Haladyna, 1982).

1. A true score is the result from administering all items in an item domain (universe).

2. An observed score is the result from administering a representative subset of a domain

3. Reliability is the squared correlation between a true and observed score.

4. An error of measurement is the difference between a true and an observed score.

5. Another type of error is one in which an examinee is incorrectly classified on the basis of an observed score.

A data base was constructed using high school student achievement test data where a unidimensional domain was established and short tests simulated to test the validity of the option weighting procedure against the standard procedure for scoring test items.

Two types of criteria were used. The first was reliability. The second was the consistency of pass-fail decisions, comparing the actual status of each examinee against the status of that examinee based on a particular test and item-scoring convention.

Procedure

Sets of 10-item, 20-item, and 30-item tests were created by randomly selecting items. For each of these nine tests, results were scored using the conventional zero-one (number right) procedure for randomly chosen 500 examinees. Each of these tests produced test scores which were then correlated with total domain scores. Each correlation involved the relationship between true (domain) and observed scores, and the squared correlation is actually reliability (Nunnally, 1978). Next, three passing standards were established, for experimental purposes. The first was the domain mean (62.4%); the second was five percentage points higher (67.4%); the third was five percentage points lower than the mean (57.4%). Then all

examinees were classified in one of two ways: (1) true pass or true fail on the bases of domain scores and the passing scores; (2) observed pass or observed fail on the basis of any of the three 10-item, 20-item, or 30-item tests. Thus it was possible to note the number of true passes and true fails correctly and incorrectly classified. A phi coefficient was computed for each relationship between true and observed pass-fail classifications as a general index of agreement--a reliability of classification consistency.

For option-weighted tests, the procedures were more involved. First all possible keyed responses were used to estimate the point-biserial correlation between each option and the total domain score. Thus all options could be evaluated first in terms of its discrimination with respect to the right option and also discrimination with respect to wrong options. Each option was weighted by its point-biserial coefficient. Research has shown that no single method is consistently better or worse, so the point biserial was chosen because of its computational simplicity and its ease of interpretation. All coefficients were subjected to a test of statistical significance ($p<.01$), to determine if a coefficient was actually zero or some non zero value. Those indexes failing to meet this criterion were treated as a zero weight in the option weighting procedure. For a test of any length therefore, the total score was simply the sum of the option weights across the items of that test.

Since some weights are negative, negative scores are possible. Also, the test scale is not in the convenient percentage correct metric for a domain-based achievement test. Therefore an algebraic transformation was done to produce domain-based scores. This conversion involved the following formula:

$$Y\% = Y + |Min|/(Max + |Min|),$$

where Y% is the transformed percentage correct score, Y is the observed score, Max is the maximum score obtainable through the weighting procedure, and Min

is the minimum score obtainable.

When all scores were transformed to Y%, correlations between the observed score and domain scores were computed as well as the percentages of correct and incorrect classifications and the corresponding phi coefficients, exactly as was done with the tests based on random selection of items.

The objective was to make the two conditions, random sampling and option weighting, as directly comparable with respect the two criterion measures of the study, reliability and consistency of classification.

Results and Discussion

Reliability Analysis

Descriptive statistics and reliability are presented in Table 1. The 10-item randomly constructed tests produced means which were consistently higher than the total domain, which is somewhat unexpected considering the random item selection. Standard deviations were also higher than expected. Reliabilities ranged from .321 to .448, rather low but not unexpected considering the shortness of these tests.

The option-weighted, 10-item test was also slightly above the domain mean, and the standard deviation was quite high. This is not surprising because option weighting has a tendency to pull examinees further apart. The reliability was .674, which is substantially higher than the reliabilities for randomly constructed tests.

With the 20-item tests, the means were closer to the domain mean of 62.4, and standard deviations were slightly lower as well. Despite the smaller standard deviations, reliabilities for randomly constructed tests were predictably higher than for corresponding 10-item tests. The option-weighted 20-item test was again significantly more reliable than the randomly selected

Table 1
Descriptive Statistics
(N=500)

Test Design	Test Length	Mean (%)	Standard Deviation (%)	Reliability
Random	10	68.5	21.4	.42
Random	10	66.5	21.7	.46
Random	10	64.4	24.0	.53
Weighted	10	63.7	25.6	.64
Random	20	62.8	20.0	.65
Random	20	63.9	21.3	.66
Random	20	64.1	19.8	.63
Weighted	20	65.4	22.5	.70
Random	30	62.4	19.8	.71
Random	30	59.1	18.4	.71
Random	30	62.4	18.9	.66
Weighted	30	64.0	21.4	.73
Total Domain	127	62.4	15.6	--

20-item tests. In fact, the option-weighted ten-item test had a higher reliability than any randomly sampled 20-item test.

With 30-item tests, means were much more closely aligned with the domain mean, as might be expected with longer tests, the standard deviations were closer to the domain standard deviation, and reliabilities followed essentially the same pattern as before. Again, a shorter 20 item option-weighted test was more reliable than any randomly sampled 30-item test. A more remarkable finding was that the option-weighted 10-item test was more reliable than any randomly selected 30-item test.

Decision Consistency

Three levels of passing scores were examined with the three different test lengths and the two scoring procedures (normal zero-one vs. option weighting). Table 2 contains information bearing on a passing score set at the domain mean of 62.4%. Table 3 presents information for all conditions where the passing score is five percentage points above the mean (67.4%). Table 4 presents information for a passing score five percentage points below the mean (57.4%). In each table and possible condition, a phi coefficient is reported as a general index of agreement. However, the percentage of persons misclassified is a more cogent statistic because one need only consider the implication of such misclassifications in certification or licensure to realize the seriousness of such errors.

Passing Score Set at the Mean. As implied in the results dealing with reliabilities, decision consistency is highest for option weighted tests for a ten-item test length. There is a sizable difference between the percentage of misclassifications for the option-weighted 10-item test and the other tests. The phi coefficient of .66 for the option weighted tests only confirms this superiority. Interestingly, errors were less frequent for true pass examinees

Table 2
Percentages of Correct and Incorrect Decisions
and
Coefficients of Agreement
Passing Score Set at the Domain Mean

Test Design	Test Length	True Pass		True Fail		Percent Mis-Classified	Phi
		Observed Pass	Observed Fail	Observed Pass	Observed Fail		
Random	10	39.8	9.0	19.0	32.0	28.0	.45
Random	10	40.2	8.6	15.8	35.4	24.4	.52
Random	10	39.8	9.0	14.0	37.2	23.0	.54
Weighted	10	42.6	6.2	11.0	40.2	17.2	.66
Random	20	42.8	6.0	12.4	38.8	18.4	.64
Random	20	43.2	5.6	11.6	39.6	17.2	.66
Random	20	42.8	6.0	14.2	37.0	20.2	.61
Weighted	20	44.2	4.6	13.0	38.2	17.6	.66
Random	30	42.6	6.2	11.0	40.2	17.2	.66
Random	30	40.0	8.8	7.8	43.4	16.6	.67
Random	30	43.4	5.4	12.4	38.8	17.8	.65
Weighted	30	44.8	4.0	9.6	41.6	11.4	.73

falsely classified as failed.

Looking at the 20-item length, an unusal exception is noted. Option weighting produced as much misclassification on the average as the randomly selected tests. The phi coefficient of .66 for option weighting is comparable to the three randomly sampled tests, .64, .66, and .61. This result can be explained by looking at the means of each test as compared with the domain mean. The 20-item option-weighted test was 3.0% easier, resulting in a shift of the distribution above the true mean. This resulted in more examinees being misclassified as pass, when they should have been failed. This problem could be resolved by equating an option-weighted test with prior forms to control for difficulty.

The 30-item option-weighted test clearly outperformed its counterparts; the margin of difference ranging from 3.0% to 4.2% percentage points, which for 500 examinees amounts of 15 to 21 examinees.

Passing Score Set 5% Above the Mean. Table 3 provides results for this analysis. In this conditiion, a fewer number of examinees will pass and a greater number will fail. In each instance, the option-weighted test performed better than randomly sampled tests. The differences were sizable for the 10-item tests and somewhat lesser for the 20-item and 30-item tests. With the longer tests, the differences seem to diminish. Option weighting seems to minimize the errors of misclassifying persons who should pass. In the 30-item test, the poorest results for option weighting were attained, again showing that with longer tests, most of the gains from option weighting appeared to be offset. Interestingly, 10-item option-weighted tests actually performed as well as 20-item randomly sampled tests, and 20-item option-weighted tests did nearly as well as 30-item randomly sampled tests.

Passing Score Set 5% Below the Mean. Table 4 provides results for this

Table 3
Percentages of Correct and Incorrect Decisions
and
Coefficients of Agreements
Passing Score Set at 5% Above the Domain Mean

Test Design	Test Length	True Pass		True Fail		Percent Mis Classified	Phi
		Observed Pass	Observed Fail	Observed Pass	Observed Fail		
Random	10	33.2	5.8	25.8	35.2	31.6	.42
Random	10	33.0	6.0	23.0	38.0	29.0	.46
Random	10	33.8	5.2	20.0	41.0	25.2	.53
Weighted	10	34.6	4.4	14.0	47.0	18.4	.64
Random	20	33.6	5.4	12.0	49.0	17.4	.65
Random	20	34.8	4.2	13.4	47.6	17.6	.66
Random	20	34.0	5.0	14.2	46.8	19.2	.63
Weighted	20	36.2	2.8	13.0	48.0	15.8	.70
Random	30	33.2	5.8	8.2	52.8	14.0	.71
Random	30	30.0	9.0	4.6	56.4	13.6	.71
Random	30	33.2	5.8	10.8	50.2	16.6	.66
Weighted	30	35.8	3.2	10.4	50.6	13.6	.73

Table 4
Percentages of Correct and Incorrect Decisions
and
Coefficients of Agreements
Passing Score Set at 5% Above the Domain Mean

Test Design	Test Length	True Pass		True Fail		Percent Mis-Classified	Phi
		Observed Pass	Observed Fail	Observed Pass	Observed Fail		
Random	10	55.2	7.4	20.4	17.0	27.8	.38
Random	10	54.6	8.0	16.2	21.2	24.2	.47
Random	10	52.4	10.2	12.6	24.8	22.8	.51
Weighted	10	52.0	10.6	7.0	30.4	17.6	.63
Random	20	56.0	6.6	9.6	27.8	16.2	.65
Random	20	53.6	9.0	10.6	26.8	19.6	.57
Random	20	54.2	8.4	12.8	24.6	21.2	.54
Weighted	20	54.6	8.0	8.2	29.2	16.2	.65
Random	30	52.4	10.2	6.8	30.6	17.0	.65
Random	30	49.4	13.2	5.4	32.0	18.6	.63
Random	30	54.4	8.2	8.2	29.2	16.4	.65
Weighted	30	55.4	7.2	7.0	30.4	14.2	.70

analysis. In this situation, a greater number of examinees pass on the basis of the lower passing score. As in the other two conditions, 10-item option-weighted tests outperformed their counterparts but with diminishing effectiveness as a function of test length. Errors of misclassification were unevenly distributed between the two types, particularly with the 30-item tests. This may be due to the fact that one randomly composed test was slightly harder than the others, and therefore more examinees failed who should have passed. Overall, these results were more definitive in comparison with the other two conditions for passing scores.

Conclusion

It has long since been established that option weighting positively affects estimates of reliability but has unclear effects on predictive validity. In this study, reliability was directly computed, and option weighting again proved effective in increasing reliability. More importantly, option weighting applied to a pass-fail decisionmaking situation provided a substantial improvement over regular scoring for tests which were constructed by simple random sampling from a well defined domain. The advantages for option weighting were greatest when the test was 10 items. As test length increased, the advantage of option weighting seemed to lessen.

While this study provides some support for the use of option weighting in decisionmaking situations, a number of questions remain unanswered. Which option weighting procedure is preferable? Point biserial weights were used, but factor analytic procedures have shown promise, and there are numerous other techniques. No study has yet comprehensively looked at all techniques.

With heightened interest in item response theory, the Rasch model or the three-parameter model might be used in conjunction with option weighting to increase score estimation at the passing score. Haladyna and Roid (1982) have

shown that selective use of items as function of difficulty will minimize error and increase domain score estimates. Thiessen (1975) used item response theory with promising results for an aptitude test, yet his study appears to be the only one where option weighting was studied in the context of item response theory.

In the present study, the tests produced by option weighting had satisfactory difficulty in most circumstances. The one exception occurred when the difficulty was discrepant from the domain mean. Had some equating been done on these scores, the results should have been more satisfactory.

A final consideration has to do with the quality of tests. If test items are well designed and field tested, all options are likely to work effectively. However, if items are not well designed, options will not work and options will not carry enough weight to make a difference and option weighting will fail. Therefore it seems reasonable to suggest that option weighting will work best in more refined testing programs and not in classroom achievement tests where items are likely to be of the poorest quality.

References

Aiken, L. R. Jr. Effects of test score variance of differential weighting on items responses. Psychological Reports, 1967, 21, 585-590.

Baker, F. B. & Hoyt, C. J. The relation of the method of reciprocal averages to Guttman's internal consistency scaling model. A paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

Bayuk, R. J. The effects of choice weights and item weights on the reliability and predictive validity of reading tests. A paper presented at the annual meeting of the American Educational Research Association, Chicago, 1974.

Bejar, I. I. & Weiss, D. J. A comparison of empirical differential option weighting scoring procedures as a function of inter-item correlation. Educational and Psychological Measurement, 1977, 37, 335-340.

Brennan, R. L. Elements of generalizability theory. Iowa City: The American College Testing Program, 1983.

Carroll, J. B. The nature of correlation, or how to choose a correlation coefficient. Psychometrika, 1961, 26, 347-372.

Claudy, J. G. Biserial weights: A new approach to test item option weighting. Applied Psychological Measurement, 1978, 2, 25-30.

Cross, L. H. & Frary, R. B. Empirical choice weighting under "guess" and "do not guess" directions. Educational and Psychological Measurement, 1978, 38, 613-620.

Davis, F. B. Estimation and use of scoring weights for each choice in a multiple-choice item. Educational and Psychological Measurement, 1959, 19, 291-298.

Davis, F. B. & Fifer, G. The effect of test reliability and validity of scoring aptitude and achievement tests with weights for every choice. Educational and Psychological Measurement, 1959, 19, 159-170.

Donlon, T. J. & Fitzpatrick, A. B. The statistical structure of multiple-choice items. A paper presented at the annual meeting of the New England Educational Research Association, 1978.

Downey, R. G. Item option weighting of achievement test items: Comparative methods of study. Applied Psychological Measurement, 1979, 3, 453-461.

Echternacht, G. Reliability and validity of item option weighting schemes. Educational and Psychological Measurement, 1976, 36, 301-309.

Guttman, L. An outline of the statistical theory of prediction. In P. Horst (Ed.) Prediction of personal adjustment. Social Science Research Bulletin, 1941, 48, 253-364.

Haladyna, T. M. An analytical and quantitative review of option-weighting research. Unpublished paper, 1984.

Hambleton, R. K., Swaminathan, H., Algina, J. & Coulsen, D. B. Criterion-referenced measurement: A review of technical issues and developments. Review of Educational Research, 78, 48, 1-47.

Hendrickson, G. F. The effect of differential option weighting on multiple choice objective tests. Journal of Educational Measurement, 1971, 8, 291-296.

Hendrickson, G. F. & Green, B. F. Jr. Comparison of factor structures of Guttman-weighted vs. rights-only-weighted tests. A paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

Henrysson, S. Gathering, analyzing and using data on testitems. In R. L. Thorndike (Ed.) Educational Measurement. Washington, D. C.: American Council on Education, 1971.

Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. New York: McGraw-Hill, 1968.

Nunnally, J. C. Psychometric Theory. New York: McGraw-Hill, 19 .

Patnaik, D. & Traub, R. E. Differential weighting by judged degree of correctness. Journal of Educational Measurement, 1973, 10, 281-286.

Raffeld, P. The effect of Guttman weights on the reliability and validity of objective tests when omissions are not differentially weighted. Journal of Educational Measurement, 1975, 12, 179-185.

Reilly, R. R. Empirical option weighting with a correction for guessing. Educational and Psychological Measurement, 1975, 35, 613-619.

Reilly, R. R. & Jackson, R. Effect of empirical option weighting on reliability and validity of an academic aptitude test. Journal of Educational Measurement, 1973, 10, 185-194.

Roid, G. H. & Haladyna, T. M. The technology of test item writing. New York: Academic Press, 1982.

Sabers, D. L. & White, G. The effects of differential option weighting of individual item responses on the predictive validity and reliability of an aptitude test. Journal of Educational Measurement, 1969, 6, 93-96

Serlin, R. C. & Kaiser, H. F. A method for increasing the reliability of a short multiple-choice test. Educational and Psychological Measurement, 1978, 38, 37-44.

Thiessen, D. M. Information in wrong responses to the Raven Progressive Matrices. Journal of Educational Measurement, 1976, 201-214.

Waters, B. The measurement of partial knowledge: A comparison between two empirical option weighting methods and rights-only scoring. Journal of Educational Research, 1975, 69, 256-260.

Willson, V. L. Maximizing reliability in multiple choice questions. Educational and Psychological Measurement, 1982, 42, 69-72.